On optimal methods of interpolation and interference phenomenon in Nonparametric Regression

Part 1

It gives me a great pleasure to present a talk at this meeting, as I was – and still consider myself – a student of Prof. Khasminskii, for more than 50 years (since 1969). Rafa, as he is affectionately addressed by all his colleagues, has always shown himself most imaginative researcher and a talented teacher, open and friendly, but not mincing his words, hospitable and charismatic, full of energy and highly experienced in teaching his students how to be successful in doing math. Much of it he must have learned in turn from his teachers A.N. Kolmogorov and E.B. Dynkin.

Needles to say that we all relish our memories from the time spent with him, including family celebrations, boating expeditions, etc. But today I'd like to emphasize what I think is the main lesson that we – his students – have learned from him. And this is that the only thing that matters in math is one's creative work, and that it should be really good!

Many other things we have learned from studying his monographs and research papers. A lesson particularly valued to me is that a good mathematics writer would always emphasize in the beginning some simple but bright idea, and then, like a magician, would derive from it a number of impressive and unexpected results.

In his first published monograph such a bright idea was the method of *Lyapunov functions*. In a later joint paper with I.A. Ibragimov (1984), a general form of the *optimal estimator* for linear functionals was derived first, from which a whole bunch of beautiful asymptotic optimality formulas could be obtained. I must admit I was returning to this last paper for decades.

So, this combination of fundamentally simple and clear ideas, which at the touch of a master can produce a whole body of impressive results, is what I've learned to appreciate most in mathematics.

For all that and much more I'm grateful to Prof. Khasminskii, and wish him to enjoy many more years of creative life, surrounded by his loving students, especially at times when it only can be done remotely.

Part II

In this presentation, I will discuss some applications of *Interpolation methods* to Nonparametric Regression. One can notice that – on one hand – such methods can be quite diverse, differing by the type of interpolating functions (polynomials, rational functions, elliptic functions, splines, etc.); various observation designs; various noise models (white noise, coloured noise); etc.

On the other hand, methods of interpolation are not generally known to be optimal, among *all* competitive approximation methods, and oftentimes may be not even appropriate, unless the data has been properly preprocessed (grouped, averaged). So, why would one focus specifically on the interpolation methods?

Well, here is the thing. The relative simplicity of interpolation methods allows one to study them in much more detail, leading to a new kind of problems and phenomena – including some meaningful non-asymptotic optimality results, – which are yet unavailable for other, more sophisticated approximation methods.

Remember what fascinated me most in the works of Prof. Khasminskii: a simple but useful idea which works well for a large variety of meaningful problems.

Now, let us consider the following standard model of random data:

$$y_j = f(x_j) + e_j, \quad j = 1, 2, \dots ,$$

where

$$\mathcal{X} = \{x_j \in \mathbf{X},\ j = 1, 2, \dots\}$$

is either a finite, or an infinite set of *observation points/nodes*, or *design*, in a given set $\mathbf{X} \subset \mathbf{R}$, and $f(x_j)$ are the values of an unknown function $f(x),\ x \in \mathbf{X}$, observed in the presence of a discrete *white noise* (not necessarily Gaussian):

$$\mathbf{E}\, e_j = 0,$$

$$\mathbf{Cov}(e_i, e_j) = \begin{cases} \sigma^2 & , \text{ if } \ i = j, \\ 0 & , \text{ if } \ i \neq j, \end{cases}$$

where σ^2 may be unknown.

Assume that f belongs to a given convex class $\mathcal{F} = \{f(x) \colon x \in \mathbf{X}\}$ of functions real valued on $\mathbf{X}$.

For a given set of *fundamental interpolating functions* $L_j(x),\ x \in \mathbf{X}$, $j = 1, 2, \dots$,

$$L_j(x_i) = \begin{cases} 1 & , \text{ if } \quad i = j, \\ 0 & , \text{ if } \quad i \neq j, \end{cases}$$

consider the following interpolating formula

$$\hat{f}(x) = \sum_j L_j(x) y_j.$$

Denote by $\mathcal{L} = \operatorname{span}\{L_j,\ j = 1, 2, \dots\}$ the linear space spanned by the fundamental interpolating functions $L_j(x)$ on $\mathbf{X}$. Call the pair $\mathcal{M} = (\mathcal{X}, \mathcal{L})$ an interpolating *design model*.

Since $\mathbf{E}y_j = f(x_j) + \mathbf{E}e_j = f(x_j)$, the expectation

$$\mathbf{E}\hat{f}(x) = \sum_j L_j(x) f(x_j) := \tilde{f}(x)$$

is just our interpolating formula applied to the true – but unknown – function $f(x)$, and $\tilde{f}(x_j) = f(x_j)$. Thus, the bias

$$b_f(x) = \mathbf{E}\hat{f}(x) - f(x) = \tilde{f}(x) - f(x)$$

is a function *oscillating* between the nodes x_j,

$$b_f(x_j) = \tilde{f}(x_j) - f(x_j) = 0, \quad j = 1, 2, \ldots .$$

Moreover, since e_j are uncorrelated,

$$\mathbf{Var}\hat{f}(x) = \sigma^2 \sum_j L_j^2(x) := \sigma^2 s(x),$$

where the *variance function*

$$s(x) = \sum_j L_j^2(x)$$

is another function *oscillating* between the same nodes,

$$s(x_j) = 1, \quad j = 1, 2, \ldots .$$

Finally, by the variance-bias decomposition, the mean squared error

$$R_f(x) := \mathbf{E}(\hat{f}(x) - f(x))^2 = \mathbf{Var}\,\hat{f}(x) + (\mathbf{E}\hat{f}(x) - f(x))^2 = \sigma^2 s(x) + b_f^2(x),$$

exhibits two *oscillating* functions, which trivially satisfy

$$\sup_{x \in \mathbf{X}} s(x) \geq 1 \quad \text{and} \quad \sup_{x \in \mathbf{X}} R_f(x) \geq \sigma^2.$$

$$\sup_{x \in \mathbf{X}} s(x) \geq 1 \quad \text{and} \quad \sup_{x \in \mathbf{X}} R_f(x) \geq \sigma^2.$$

These inequalities lead us to two obvious criteria of optimality.

Definition 1. A design model $\mathcal{M} = (\mathcal{X}, \mathcal{L})$ is ***D*-optimal**, if

$$\sup_{x \in \mathbf{X}} s(x) = 1.$$

Definition 2. A design model $\mathcal{M} = (\mathcal{X}, \mathcal{L})$ is ***R*-optimal**, with respect to a given functional class $\mathcal{F}$, if for all $f \in \mathcal{F}$,

$$\sup_{x \in \mathbf{X}} R_f(x) = \sigma^2.$$

Let us try to understand better these two definitions.

Generally speaking, D-optimal models exist, but are rare; e.g., according to the *Optimal Design Theory*, for any finite-dimensional space $\mathcal{L}$ on $\mathbf{X}$, there exists, typically, exactly one design $\mathcal{X} \subset \mathbf{X}$ (oftentimes, far from obvious), such that the model $\mathcal{M} = (\mathcal{X}, \mathcal{L})$ is D-optimal.

Now, let us assume that a model $\mathcal{M} = (\mathcal{X}, \mathcal{L})$ is D-optimal, i.e.,

$$\sup_{x \in \mathbf{X}} \sigma^2 s(x) = \sigma^2.$$

Then, to be R-optimal, it should satisfy additionally

$$\sup_{f \in \mathcal{F}} \sup_{x \in \mathbf{X}} R_f(x) = \sup_{x \in \mathbf{X}} (\sigma^2 s(x) + \sup_{f \in \mathcal{F}} b_f^2(x)) = \sigma^2 \quad !$$

How can this be? Obviously, this may happen only when there is some kind of interaction between the two oscillating functions: the variance and the squared bias. In physics, such kind of interaction is called *interference*, in which two oscillating functions may mutually dampen, or even completely eliminate, one another.

More precisely, denote

$$\sup_{f\in\mathcal{F}} b_f^2(x) = (f^\circ(x))^2, \tag{1}$$

and assume that for some $c > 0$,

$$s(x) \le 1 - c(f^\circ(x))^2. \tag{2}$$

Then, according to (1)–(2), one observes the following *interference* effect

$$\sup_{f\in\mathcal{F}}\sup_{x\in\mathbf{X}} \mathbf{E}(\hat f(x) - f(x))^2 = \sup_{f\in\mathcal{F}}\sup_{x\in\mathbf{X}} \left(\sigma^2 s(x) + b_f^2(x)\right) \le$$

$$\sigma^2(1 - c(f^\circ(x))^2) + (f^\circ(x))^2 = \sigma^2 + (1-\sigma^2 c)(f^\circ(x))^2 \le \sigma^2,$$

provided that

$$\sigma^2 \ge \sigma_0^2 = \frac{1}{c}.$$

Under these assumptions – and essentially only in this case – a *D*-optimal design turns out to be simultaneously *R*-optimal.

This brings us to the following two questions:

1. How to find $f^\circ(x)$ in (1)?
2. How to find c in (2)?
Remark. Obviously σ_0^2 cannot be 0, but it can be arbitrary small, and – from the interference point of view – the smaller, the better.

It turns out that for some typical functional classes $\mathcal{F}$ (*Sobolev, Hardy*), the *Optimal Recovery Theory* may allow one to determine simultaneously **i)** the *D*-optimal design model $(\mathcal{X}, \mathcal{L})$; and **ii)** the function $f^\circ(x)$ in Question **1,** all expressed in terms of the so-called *extremal element* $f^\circ(x)$, which is the solution of the following extremal problem:

$$\sup_{f\in\mathcal{F}:\ f(x_j)=0,\ j=1,2,\dots} |f(x)|.$$

Finding the value of c in Question **2** is a separate problem. Let me next illustrate these two problems by some examples.

Example 1. Let $\mathcal{F} = H(r, Q)$ be the Hardy class of 1-periodic functions $f(x)$, real valued on R, and admitting a bounded analytic continuation into the strip

$$S_r = \{z = (x + iy) : |y| < r\} \subset \mathbb{C},$$

such that

$$\sup_{z \in S_r} |f(z)| \leq Q.$$

The optimal n points design $\mathcal{X}$ in $\mathbf{X} = [0, 1)$ is any shift of the nodes

$$x_j = \frac{j}{n}, \quad j = 0, ..., n-1.$$

The problem of recovering function $f \in \mathcal{F}$ from its values $f(x_j)$ is an *Optimal Recovery problem*, with the following *extremal element*. Let

$$\mathrm{sn}\,(x; k)$$

be the *Jacobi elliptic function* of *modulus* k (an oscillating function called "elliptic sinus"). This is a well known doubly periodic analytic function in $\mathbb{C}$, $\mathrm{sn}\,(0; k) = 0$, which has real period denoted $4\mathbf{K}$, pure imaginary period denoted $2i\mathbf{K}'$, and a simple pole it $i\mathbf{K}'$, all determined by the parameter k. The modulus k is to be chosen so that

$$\frac{\mathbf{K}'}{\mathbf{K}} = 4r.$$

If $k = 0$, $\mathrm{sn}\,(x; 0) \equiv \sin x$, whereby $\mathbf{K} = \pi/2$, $\mathbf{K}' = \infty$. There is also an "elliptic cosine" denoted $\mathrm{cn}\,(x; k)$ such that

$$\mathrm{cn}^2(x; k) + \mathrm{sn}^2(x; k) = 1.$$

The problem of recovering an unknown function $f \in \mathcal{F}$ from its values $f(x_j)$ has the following *extremal element* oscillating at the nodes x_j:

$$f^\circ(x) = Q \prod_{j=0}^{n-1} \mathrm{sn}\,(2\mathbf{K}(x - x_j); k).$$

The optimal method of recovery is then given by the interpolating formula

$$\hat{f}(x) = \sum_{j=0}^{n-1} L(x - x_j) y_j,$$

where the *interpolating kernel* $L(x)$ is defined as

$$L(x) = 2\mathbf{K} \frac{f^{\circ}(x)}{f^{\circ\prime}(0)} \frac{\mathrm{cn}\,(2\mathbf{K}x)}{\mathrm{sn}\,(2\mathbf{K}x)}$$

if n is even, and slightly differently if n is odd. In the trigonometric case $k = 0$, $L(x)$ would coincide with the so-called *modified Dirichlet kernel,*

$$L(x) = \frac{\sin \frac{\pi n x}{2}}{\tan \frac{\pi x}{2}},$$

for even n (or with the standard Dirichlet kernel, for n odd). Then, for any $f \in \mathcal{F} = H(r, Q)$,

$$|b_f(x)| = |\tilde{f}(x) - f(x)| \le |f^{\circ}(x)|,$$

and this bound in optimal, for every $x \in [0, 1]$.

The mean squared error of the interpolant $\hat{f}(x)$ satisfies

$$\mathbf{E}(\hat{f}(x) - f(x))^2 = \sigma^2 s(x) + b_f^2(x) \le$$

$$\sigma^2(1 - c(f^{\circ}(x))^2) + (f^{\circ}(x))^2 = \sigma^2 + (1 - c\sigma^2)(f^{\circ}(x))^2 \le \sigma^2,$$

if

$$\sigma^2 \ge \sigma_0^2 = \frac{1}{c} = 4e^{-\pi n r} + O(e^{-3\pi n r}), \quad n \to \infty.$$

The value of c can be determined through a comparison of the variance function $s(x)$ and $(f^{\circ}(x))^2$, at the pole $i\mathbf{K}'$; for more details and further references see BL(2016).

Example 2. Let $\mathbf{X} = \mathbf{R}$ and $\mathcal{X} = \mathbb{Z} = \{j : j = 0, \pm 1, \pm 2, ...\}$. Consider again the Hardy class $\mathcal{F} = H(r, Q)$ of functions f, real valued on $\mathbf{X}$ and bounded by Q in the strip S_r (now not necessarily periodic).

If we transformed the strip S_r conformally into the interior of the unit disk $U = \{z : |z| < 1\}$, by the mapping

$$z = \tanh \frac{\pi w}{4r}, \quad w \in S_r, \quad z \in U,$$

so that the nodes $j \in \mathbb{Z}$ were mapped into some nodes $z_j \in (-1, 1)$, then the corresponding extremal element $f^\circ(z),\ z \in U$, would be the infinite *Blaschke product*

$$W(z) = Q \prod_{j=-\infty}^{\infty} \mathrm{sgn} z_j \frac{z - z_j}{1 - z_j z}.$$

If then we mapped the unit disk U back into the strip S_r, $W(z)$ would transform into the following extremal element in $\mathcal{F} = H(r, Q)$,

$$f^\circ(x) = Q\sqrt{k}\, \mathrm{sn}\,(2\mathbf{K}x; k),$$

where the modulus k again is to be chosen so that

$$\frac{\mathbf{K}'}{\mathbf{K}} = 4r.$$

The corresponding optimal interpolating kernel is

$$L(x) = \frac{\pi}{2r\mathbf{K}} \frac{\mathrm{sn}\,(2\mathbf{K}x; k)}{\sinh \frac{\pi x}{2r}}.$$

The interpolating formula

$$\hat{f}(x) = \sum_{j=-\infty}^{\infty} L(x-j)y_j, \tag{3}$$

satisfies for all $f \in \mathcal{F}$

$$|b_f(x)| = |\tilde{f}(x) - f(x)| \le |f^\circ(x)|,$$

and this bound is optimal, for every x. Its mean squared error is

$$\mathbf{E}(\hat{f}(x) - f(x))^2 = \sigma^2 s(x) + b^2(x) \le$$
$$\sigma^2(1 - c(f^\circ(x))^2) + (f^\circ(x))^2 = \sigma^2 + (1 - c\sigma^2)(f^\circ(x))^2 \le \sigma^2,$$

if

$$\sigma^2 \ge \sigma_0^2 = \frac{1}{c},$$

where

$$\sigma_0^2 = \frac{1}{c} \le 2\sqrt{6}Q^2 r e^{-2\pi r}(1 + O(r^{-1})), \quad r \to \infty.$$

The constant c has been determined through a comparison of the essential terms of Laurent expansions for $s(x)$ and $(f^\circ(x))^2$, at the pole $i\mathbf{K}'$.

In the limiting case $r \to \infty$, the interpolating formula (3) becomes the famous Shannon-Nyquist-Kotel'nikov *cardinal series*

$$\hat{f}(x) = \sum_{j=-\infty}^{\infty} \frac{\sin \pi(x-j)}{\pi(x-j)} y_j,$$

which is an unbiased estimator for the class $\mathcal{E}_\pi$ of band-limited functions whose spectrum is limited to $[-\pi, \pi]$. In this exceptional case, $s(x) \equiv 1$ and $b_f(x) \equiv 0,\ f \in \mathcal{E}_\pi$, so that $\sigma_0^2 = 0$ and $\hat{f}(x)$ is *trivially* both D-optimal and R-optimal with respect to $\mathcal{E}_\pi$. For more details, see BL(2018).

Example 3. Cardinal spline interpolation. Let again $\mathbf{X} = \mathbf{R}$ and $\mathcal{X} = \mathbb{Z}$. Consider the interpolating formula

$$\hat{f}_m(x) = \sum_{j=-\infty}^{\infty} L_m(x-j) y_j ,$$

where the fundamental interpolating kernel $L_m(x)$,

$$L_m(j) = \begin{cases} 1 & , \quad \text{if} \quad j = 0, \\ 0 & , \quad \text{if} \quad j \neq 0, \end{cases}$$

is an m-th degree *cardinal spline*, i.e., $L_m(x) \in \mathbf{C}^{(m-1)}(\mathbf{R})$, and $L_m(x)$ coincides with some m-th degree polynomial, on each interval

$$\begin{cases} (i, i+1) & , \quad \text{if} \quad m \text{ is even}, \\ (i-1/2, i+1/2) & , \quad \text{if} \quad m \text{ is odd}. \end{cases}$$

Such an interpolating kernel $L_m(x)$ exists, is unique, and decreases exponentially when $x \to \pm\infty$. It can be represented as an infinite series of the shifts $B_m(x-j)$ of B-splines $B_m(x)$ of the same degree m.

Now, consider the Sobolev space

$$\mathcal{F}_m = S(m, C, \varrho) = \{ f : |f^m(x)| \leq C m! \varrho^m \} .$$

Here ϱ can be interpreted as a smoothness parameter.

With respect to the bias $b_f(x) = \tilde{f}_m(x) - f(x)$, the interpolant $\hat{f}_m(x)$ is optimal in $\mathcal{F}_{m+1}$, at every given point x. The extremal element in $\mathcal{F}_{m+1}$ is proportional to the so-called *perfect 2-periodic Euler spline* $S_{m+1}(x)$ of degree $m+1$, having zeros at all knots $j \in \mathbb{Z}$.

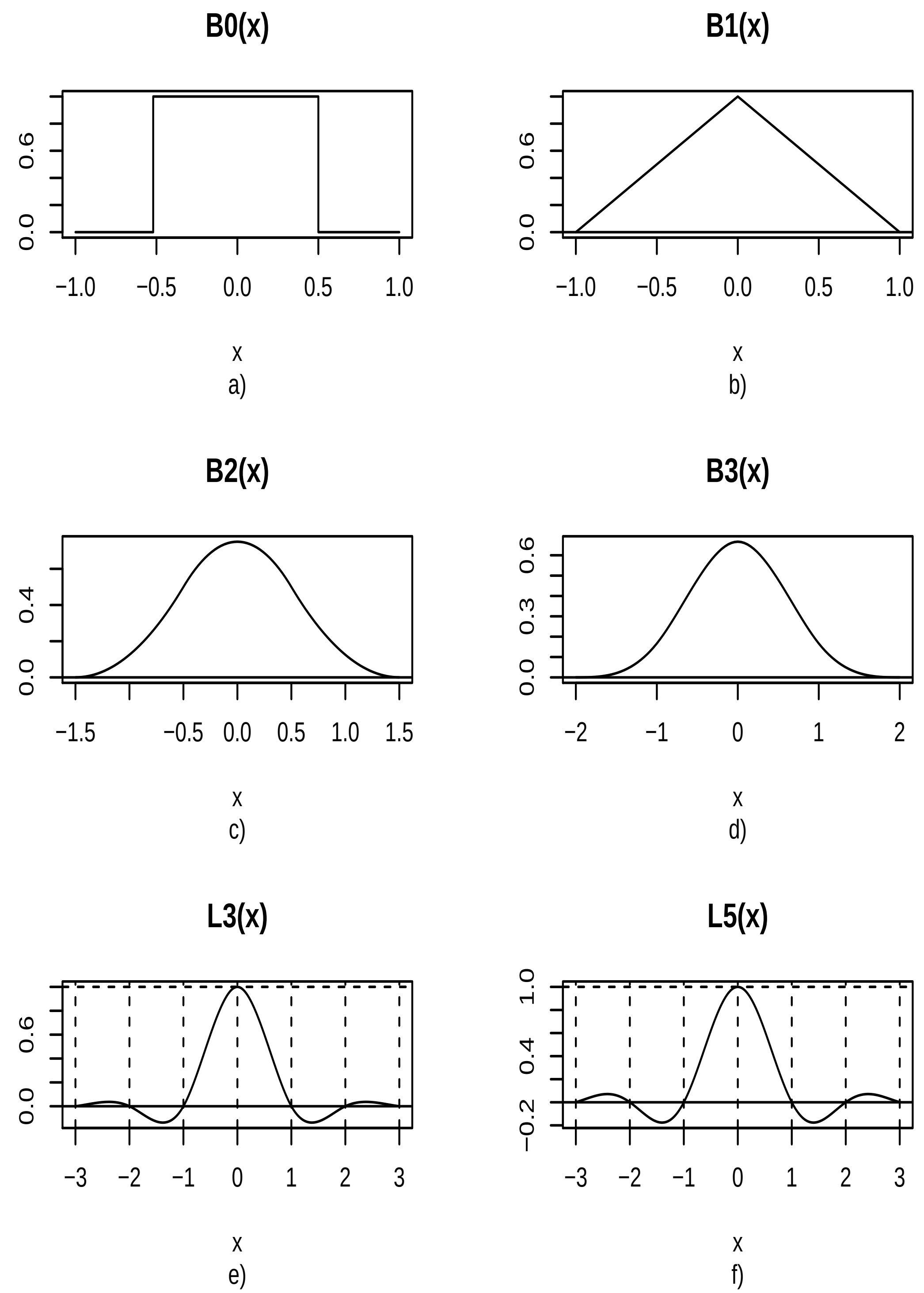

Figure 1: *B*-splines of degree m: a) m = 0; b) m = 1; c) m = 2; d) m = 3; and interpolating kernels of degree m: e) m = 3; f) m = 5

The *perfect Euler splines* $S_n(x)$ can be conveniently represented in the form of Fourier series,

$$S_n(x) = \begin{cases} \frac{4(-1)^{\lfloor \frac{n}{2} \rfloor}}{\pi^{n+1}} \sum_{j=0}^{\infty} \frac{\sin(2j+1)\pi x}{(2j+1)^{n+1}} & , \quad \text{if } n \geq 0 \text{ is even,} \\ \frac{4(-1)^{\lfloor \frac{n}{2} \rfloor}}{\pi^{n+1}} \sum_{j=0}^{\infty} (-1)^j \frac{\sin(2j+1)\pi x}{(2j+1)^{n+1}} & , \quad \text{if } n > 0 \text{ is odd.} \end{cases}$$

The fact that $S_n^\circ(i) = 0$ for all $i \in \mathbb{Z}$ and is 2-periodic is obvious. The fact that $S_n^\circ(x)$ is an n-th degree perfect spline is easier to see for n even, since in that case, by a textbook example of Fourier series,

$$(S_n(x))^{(n)} = \frac{4}{\pi} \sum_{j=0}^{\infty} \frac{\sin(2j+1)\pi x}{2j+1} = \begin{cases} 1 & , \quad 0 < x < 1, \\ -1 & , \quad 1 < x < 2, \\ 0 & , \quad x = 0, 1, 2. \end{cases}$$

A similar argument can be used for n odd.

These splines are also related to the classical *Euler polynomials*. For n even, they coincide – up to a constant – with the corresponding Euler polynomials on $[0, 1]$:

$$S_n(x) = \frac{1}{n!} \begin{cases} x^2 - x & , \quad n = 2, \\ x^4 - 2x^3 + x & , \quad n = 4, \\ x^6 - 3x^5 + 5x^3 - 3x & , \quad n = 6, \\ \text{etc.} \end{cases}$$

For n odd, there is a similar – but different – relation, since for such n, Euler polynomials have zeros at half-integer knots $i + 1/2$.

The fact that $S_m(x)$ is proportional to the extremal element implies that for all $f \in \mathcal{F}_{m+1}$, the bias of our interpolant satisfies

$$|b_f(x)| \leq C(m+1)!\varrho^m \left| S_{m+1}(x) \right|,$$

and this bound is unimprovable.

Now, to find the corresponding variance function,

$$s_m(x) = \sum_{j=-\infty}^{\infty} L_m^2(x-j),$$

consider, following Schoenberg (1973), the so-called m-th degree *exponential Euler spline*

$$S_m(x;z) = S_m(x;e^{iu}) = \sum_{j=-\infty}^{\infty} L_m(x-j)e^{iju},$$

where $u \in [0, 2\pi]$ and

$$x \in \begin{cases} [0,1] & , \text{ if } \ m \text{ is even,} \\ [-1/2, 1/2] & , \text{ if } \ m \text{ is odd.} \end{cases}$$

Then, by the *Parseval identity*,

$$s_m(x) = \sum_{j=-\infty}^{\infty} L_m^2(x-j) = \frac{1}{2\pi}\int_0^{2\pi} |S(x,e^{iu})|^2\, du.$$

Remark. An explicit determination of the exponential Euler splines $S_m(x;z)$ can be, probably, viewed as the top achievement of Schoenberg's spline theory. Function $S_m(x;z)$ in it appears to be an m-th degree polynomial in x, whose coefficients $a_i = a_i(z)$ are explicitly expressed in terms of the so-called *Euler-Frobenius polynomials*.

Calculation of the variance function $s_m(x)$ is trivial for $m = 0, 1$ (piecewise constant or linear splines); reduces to some tabulated integrals for $m = 1, 2$ (quadratic or cubic splines); and rapidly becomes quite challenging already for $m = 4, 5$ (quartic or quintic splines), and beyond. (In particular, it requires some use of Maple). It turns out that for all m, the cardinal spline interpolating formula is D-optimal, i.e.,

$$\sup_{x \in \mathbf{X}} s_m(x) = 1,$$

although the proof for arbitrary $m > 1$ is extremely difficult.

Let us now summarize the result concerning R-optimality for the Sobolev class $\mathcal{F}_{m+1} = S(m+1, C, \varrho) \quad (m = 1, ..., 5)$, where ϱ can be perceived as a *smoothness parameter*. Let us provisionally divide all Sobolev classes into those with $0 < \varrho < 1$ ("smooth" Sobolev classes), and those with $\varrho \geq 1$ ("non-smooth" Sobolev classes). The interpolant $\hat{f}_m(x)$ is R-optimal with respect to $\mathcal{F}_{m+1}$, if $\sigma^2 \geq \sigma_0^2$, where

$$\sigma_0^2 = c_m C^2 \varrho^{2(m+1)},$$

and

$$0 < c_m \leq 5.5 \quad \text{for} \quad m = 1, ..., 5.$$

Obviously, from the interference' point of view, the smaller is σ_0^2, the better. Thus, the criterion leads to the following conclusions:

For "smooth" Sobolev classes (with $0 < \varrho < 1$), the higher is degree of cardinal spline interpolants, the better (σ_0^2 is smaller).

For "non-smooth" Sobolev classes (with $\varrho > 1$), the lower degree cardinal spline interpolants are better (σ_0^2 is smaller).

Of course, the smoothness parameter ϱ is never known precisely. This invites some adaptive approach à la Lepski.

Another problem, yet to be addressed, is finding optimal (but not R-optimal) interpolants, when $\sigma^2 < \sigma_0^2$, where only partial interference may occur.

THANK YOU!